Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

REINFORCEMENT LEARNING IN A DIGITAL TWIN FOR GALVANO HOIST SCHEDULING

Marvin Carl May[1,2], Louis Schäfer[3,4], and Jan-Philipp Kaiser[3]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA,
UNITED STATES OF AMERICA
[2]School of Mechanical & Aerospace Engineering, Nanyang Technological University, SINGAPORE
[3]wbk Institute of Production Science, Karlsruhe Institute of Technology, Karlsruhe, GERMANY
[4]adesso SE, Dortmund, GERMANY

ABSTRACT

Reinforcement Learning (RL) has evolved as a dominant AI method to move towards optimal control of complex systems. In the domain of manufacturing, production control has emerged as one of the major application areas, where material flow is governed by an RL agent that is fed with the real-time information flow of the system through a digital twin. A digital twin framework facilitates efficient and effective RL training. The coordination task performed in discrete, flexible manufacturing offers multiple decisions that cannot be found in all cases. Hoist scheduling for galvanic equipment introduces additional constraints as parts cannot be left inside a galvanic bath arbitrarily long. Even short deviations critically affect product quality, which is even more complicated in high mix high volume environments. The proposed RL agent learns superior control compared to the state-of-the-art and simple heuristic rules can be derived for everyday application in the absence of digital twins.

1 INTRODUCTION

In manufacturing simulation is the foundation of product, process and system design. Mechanical and electrical properties are simulated to guarantee certain requirements and behavior. Process simulations are often used to identify quality deficiencies or improve equipment. On an operational level simulations have become the de facto standard tool to analyze the dynamical behavior of larger systems to improve their design and operations. With an interconnection to the real system, its decisions and data streams a digital twin emerges (Lugaresi and Matta 2018). However, the point of focus typically remains on a single level, in our domain operations and system simulations, often in form of event discrete or agent based simulations, coupled with digital twins with a focus on increasingly flexible manufacturing (Overbeck et al. 2024). Recognizing the effects of operational decisions on instance level product quality is often hard to measure and as a result often neglected (May et al. 2024). Galvanic equipment offers a rare chance to optimize holistically, as the time products spend in different baths critically influences their quality on the process side and the scheduling of the hoist, to take products out of baths and bring them in, which heavily influences the operational performance of such a flexible system.

From a manufacturing principles perspective, a flexible production system can be conceptualized as job-shop style production at the shop floor level, attributed to the high flexibility of the conveyor system (May et al. 2023). Alternatively, the flexible production line can be categorized as a hybrid flow principle. While specialized workstations are sequentially arranged according to the work process, the material flow is not necessarily unidirectional. For instance, utilizing a workstation as both input and output store results in reversed material flow within the production line, creating a hybrid flow shop arrangement. Production control complexity in flexible production lines is further exacerbated by the necessity to monitor maximum dwell times of intermediate products at individual workstations. Products exceeding the maximum allowable dwell time at certain stations before transport to subsequent stations are considered damaged. Some stations

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

may impose a permissible dwell time of 0 seconds, necessitating immediate transport of the order (Paul et al. 2007).

The dimensional parameters of flexible production lines can influence workstation arrangements, potentially eliminating buffer stations, showcasing the need for integrated design and operations optimization (Manier and Lamrous 2008). Consequently, intermediate products must be transported directly to the next downstream station for subsequent processing. If the next station's current job remains unreleased, the product is retained at the current workstation, potentially exceeding critical transport time windows (May et al. 2021). In electroplating applications, the production environment constitutes a flexible production line where raw components undergo transformation into semi-finished or finished products through sequential processing steps (Manier and Bloch 2003). Unlike conventional manufacturing setups, the workstations in electroplating plants are not machines but rather baths that perform electrochemical or thermal treatments through immersion processes. Components are transported between baths in predetermined sequences by one or more conveyors (Paul et al. 2007). Figure 1 illustrates an electroplating plant configuration featuring two conveyors and a single carrier (in the front) functioning as both the loading and unloading point for production operations. Hoist scheduling in this context refers to the control problem of moving hoists to take products out of baths at the right time to avoid over- or under-treatment, transport them to the next bath or carrier loader or unloader and move to the next position (Manier and Bloch 2003). Research extensively studied optimization based scheduling approaches, however the reality in larger (>1 hoist) systems remain traditional heuristic rules due to frequent uncertain events, outdated schedules, high complexity and operators' preferences (Reimschüssel et al. 2023).

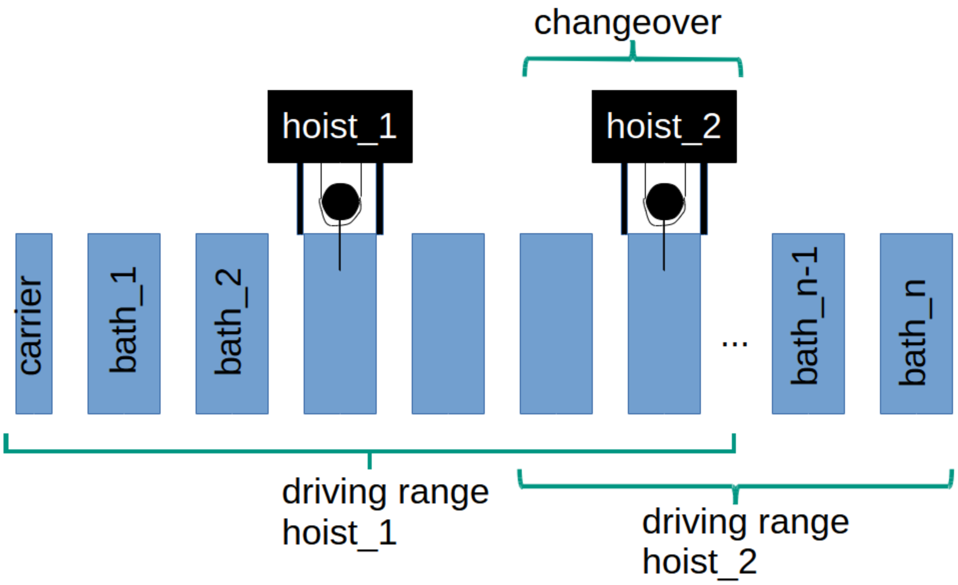

Figure 1: Visualization of the hoist scheduling in electroplating with exemplary two hoists and a carrier loader that serves as material flow entry and exit.

Electroplating in general constitutes a fundamental galvanic manufacturing technology utilized for surface property modification of physical components (Paul et al. 2007). The underlying electrochemical process facilitates the deposition of metallic layers on substrate surfaces through direct current application in an electrolytic cell, where the component functions as the cathode while the metal to be deposited serves as the anode within a metal salt solution (Subaï et al. 2006). The process enables precise manipulation of surface characteristics, enhancing corrosion resistance, wear properties, electrical conductivity, optical reflectivity, and aesthetic qualities (Manier and Bloch 2003). Common electroplating metals include zinc, nickel, chromium, copper, gold, and silver, each conferring specific functional attributes to the substrate.

Deposition thickness can be regulated through parameters including current density, temperature, bath composition, and immersion duration and besides product quality also affects the sustainability of the process (Subaï et al. 2006). Industrial applications of electroplating can encompass protective coatings against environmental degradation, enhancement of electrical and mechanical properties, and provision of decorative finishes across numerous sectors including automotive, aerospace, electronics, and consumer goods. The latter is increasingly prevalent in modern manufacturing, for instance to make plastic parts in interior design with a metal touch, thereby comprising multiple galvanic process with stringent requirements (Reimschüssel et al. 2023). The technique permits selective application, offering significant manufacturing versatility.

Thus, real-time control, enabled by a digital twin, and brought forward by intelligent control algorithms is a challenge for the galvano electroplating industry. For such a flexible manufacturing control problem, the application of reinforcement learning as an intelligent machine learning algorithm seems promising (Chen et al. 2023). This paper, thus, provides a digital twin and reinforcement learning based approach for hoist scheduling.

2 LITERATURE REVIEW

The control and optimization of conveying equipment in electroplating facilities presents significant challenges that vary considerably depending on the specific production environment. To maximize system performance metrics such as throughput or resource utilization, numerous system characteristics must be accommodated, including flexible processing times of baths, non-negligible transport times, resource capacities (baths, hoists, carriers), maximum dwell time constraints, and processing sequence requirements. Numerous research efforts have concentrated on single-hoist scenarios within the reactive hoist scheduling problem (RHSP) domain. Yih and Thesen (1991) applied expert knowledge to identify states and transition probabilities, subsequently transforming the problem into a semi-Markov optimization framework for an electroplating facility with one hoist, five baths, and heterogeneous job families. (Yih 1992) further explored this approach by developing a methodology where planning rules are extracted from optimal strategies derived through semi-Markov decision models based on collected planning data. Building upon this foundation, Yih et al. (1993) proposed a hybrid system integrating human knowledge, semi-Markov decision modeling, and artificial neural networks.

Heuristic methodologies have also been developed for single-hoist RHSPs. Chauvet et al. (2000) introduced the Forward-Backward-Earliest-Start-Time algorithm (FBEST), while Chové et al. (2009) presented an algorithm based on various heuristic rules governing bath and hoist behavior. Addressing more complex scenarios, Chtourou et al. (2013) proposed a heuristic algorithm for cyclic hoist scheduling with two hoists sharing a common track, where collision avoidance is critical. Their approach initially generates a set of sequences, assigns moves to hoists for each sequence, and employs Mixed Integer Linear Programming (MILP) to determine optimal starting times for hoist movements.

Recent literature has introduced more sophisticated approaches to hoist scheduling optimization. Chen et al. (2024) developed a Task sequence-Hoist scheduling Coupled Optimization (THCO) model that simultaneously addresses task sequence and hoist scheduling requirements, employing an Improved Salp Swarm Algorithm (ISSA) with enhanced convergence properties. Their research demonstrated that coupled optimization produces superior production schemes compared to approaches that consider these elements separately. Ptuskin et al. (2024) investigated cyclic multi-hoist scheduling with fuzzy processing times, introducing a mathematical model utilizing fuzzy numbers to address complex galvanic lines served by multiple hoists. Their work represents the first application of fuzzy set theory to multi-hoist scheduling problems, extending the "method of prohibited intervals" previously limited to single-hoist scenarios. For large-scale applications, Xiao et al. (2024) formulated electroplating scheduling as a temporal planning problem using adapted PDDL (Planning Domain Definition Language) and developed a hierarchical temporal planning approach that efficiently generates high-quality solutions for real-life benchmark instances.

Lee and Kim (2022) explored reinforcement learning applications for robotic flow shop scheduling with processing time variations, related to the hoist scheduling problem. Their model demonstrated superior performance compared to traditional first-in-first-out (FIFO) and reverse sequence (RS) approaches, highlighting the applicability of reinforcement learning techniques to robotic flow shop scheduling problems. In a similar vein, reinforcement learning in general has become increasingly popular for dispatching problems (Overbeck et al. 2021). Increasing robustness and generalizability of RL implementations in scheduling pave the way for their application to hoist scheduling (Overbeck et al. 2023).

Reimschüssel et al. (2023) addressed the automation vendor perspective in hoist scheduling, identifying that while a broad spectrum of electroplating scheduling problems can be modeled as mixed integer programs, real-world implementations require automated solutions with reduced engineering effort. They proposed combining mixed integer programming with derived hyper-heuristics to bridge this research gap. This need to derive heuristics from the learned reinforcement learning policies is universally present and can be addressed with fitting of a decision tree to the decision policy (Kuhnle et al. 2022). Most electroplating companies are in the small and medium enterprise (SME) domain and can benefit from such machine learning and artificial intelligence use-cases (Beiner et al. 2023).

Based on these findings the first research question can be derived: Can the reactive hoist scheduling problem be solved by reinforcement learning? Reinforcement learning is an ideal fit to the problem at hands, has not been studied yet but also offers advantages beyond pure performance. The underlying requirement for RL however is the presence of a cyber-physical system as a prerequisite for training and deployment. Ideally a digital twin is applied.

Leiden et al. (2021) adopted a cyber-physical production system approach, integrating energy and resource flow simulation to create a digital twin of physical plating lines. Their implementation at an industrial acid zinc-nickel plating facility demonstrated potential electricity and resource savings of up to 10% through enhanced process transparency and scenario simulation. May et al. (2021) prepare a digital twin for dispatching and reinforcement learning purposes in a similar, flexible manufacturing environment. The approach however does not regard product quality and deviations. In semiconductor manufacturing such a digital twin can be realized (May et al. 2024) and even automatically created (May et al. 2024). In electroplating however, such an advanced system is still missing. Thus, the second research question can be deduced as: what must be contained in a digital twin for optimized production control in electroplating?

3 DIGITAL TWIN IN ELECTROPLATING PRODUCTION SYSTEMS

A digital twin for an electroplating production system can be conceptualized as a discrete event simulation (DES) that transforms rigid heuristic rule- and intuition-based control process decisions into data-driven decision-making frameworks. The digital twin concept extends beyond simple simulation to create a synchronized virtual representation capable of real-time system evaluation and prediction (Lugaresi and Matta 2018). In the context of electroplating systems with multiple hoists and baths, this approach offers significant advantages for operational optimization, in a similar vein to a foresighted digital twin (May et al. 2021). Due to the high volume high variance environment (Reimschüssel et al. 2023), using a digital twin in electroplating for operations offers value, that in this study is to be realized through the reinforcement learning application.

3.1 Digital Twin Architecture for Electroplating Systems

The electroplating system under consideration, simplified visualized in Figure 1, presents substantial complexity with two hoists operating on a shared track, 23 baths and positions (three of which facilitate transfer between hoists), and multiple stochastic events affecting system performance. These stochastic elements include equipment failures (baths, hoists, material holders), variable loading/unloading durations, manual interventions, and hoist movements to avoid collisions.

To effectively model this complexity, the python based DES software Simpy provides an appropriate foundation for the digital twin and for the implementation of reinforcement learning (Loffredo et al. 2024; Wurster et al. 2022). This approach enables flexible representation of the system's numerous constraints and operational parameters, including (Manier and Bloch 2003; Chen et al. 2024; Ptuskin et al. 2024): (1) Flexible processing times that vary by product type, material, size, and processing conditions, (2) Non-negligible transport times that impact productivity and quality, (3) Resource capacity limitations across baths, hoists, and carriers, (4) Maximum and minimum dwell time constraints for products left in baths that can result in product damage if exceeded and (5) Processing sequence requirements specific to each order type. In order to preserve product quality, exceeding maximum dwell time is not permissible as in certain cases it requires rework or dwell time adaptions in subsequent baths or may require scrapping.

3.2 Process Characteristics and Modeling Challenges

Electroplating operations present unique modeling challenges for simulations and digital twins (Leiden et al. 2021). Each rack represents a single order without batching capabilities, and each order follows a fixed procedural program specifying the required bath types and sequence. These programs include minimum and maximum duration (dwell time) parameters for each bath, with high criticality assigned to baths where the difference between these times is minimal. As maximum duration approaches, order criticality increases proportionally.

The system operates as a one-piece-flow, allowing only a single order at a time in each hoist or bath. This characteristic, combined with the absence of buffer capacity in baths due to maximum dwell time constraints, creates a tightly coupled system where timing is critical (Paul et al. 2007). Processing times for each order-bath combination fall within defined t_{min} and t_{max} parameters, with durations outside these bounds resulting in defective products or largely disrupted schedules (Reimschüssel et al. 2023). The significant variation in available time windows between different bath types introduces substantial scheduling complexity. Further complicating the system is the presence of alternative (duplicate) baths and time-coupled processing sequences. The single-track constraint for hoist movement introduces additional operational challenges, including collision risks when hoists attempt to reach the same position simultaneously and access limitations where certain baths can be exclusively reached by a single hoist (Manier and Lamrous 2008). These factors create potential deadlock scenarios where a hoist may be unable to reach a necessary order. This can be incorporated into the digital twin and later the RL framework. For the sake of simplicity and with respect to the regarded industrial case study, galvanic cross converters that move object between tracks, in larger electroplating systems, and their control are not regarded in this study.

3.3 Digital Twin Implementation

Constructing the digital representation requires fitting statistical distributions to observed historical data (May et al. 2024). While data availability may impose limitations on observation periods, sufficient data is necessary to recreate the current status of orders, equipment, and their interrelationships. The latter can be obtained from basic IT systems currently used to monitor the bath, carrier or hoist status and verify the digital twin as a real twin to the underlying system through a simplified online validation (Lugaresi et al. 2022). Missing data, particularly regarding ongoing processes, can be imputed based on historical observations and derived statistical distributions (May et al. 2024). The DES structure provides a flexible framework capable of accommodating the system's variability while maintaining the logical relationships between components and constraints that define the electroplating operation.

The regarded use case specifically has 13 carriers in 23 baths plus additionally 3 storage spaces and a carrier loader or unloader as specified in Figure 2. The number of carriers is crucial as it is the maximum number of products that can be operated at the same time. New products and their jobs enter the system continuously and include priority jobs. Besides uncertain loading and unloading times, the time to release and fetch carriers from and in baths similarly varies, depending on the bath, product type and other

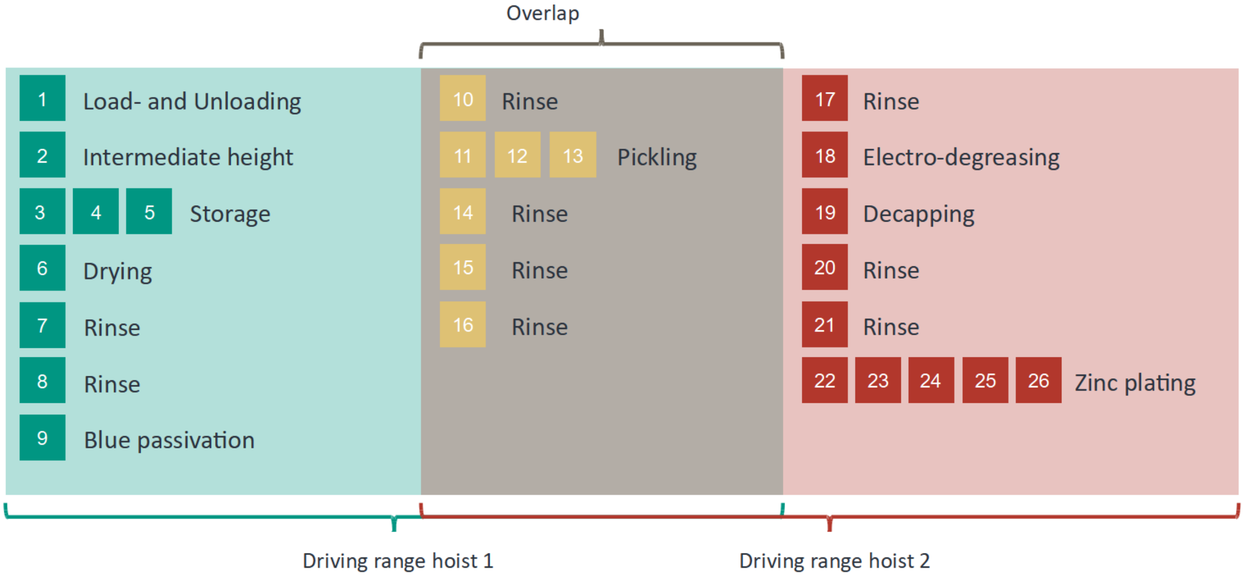

Figure 2: Use case visualization.

influencing factors. The maximum length of a job in this case can be considered to not exceed 15 baths to be visited, including, cleaning and drying.

3.4 Digital Twin Instantiation

Through an every minute query of the primitive system IT, the current status of the system and its components is fed into the digital twin. At the same time, the statistics of the operation and performance are added to the storage and knowledge base to enable foresight. Manual data entries are made for the bath type and have to be updated during changes. Throughout the study time no baths were switched, for the benefit of the digital twin, but also for the benefit of the trained RL agent. The digital twin is encapsulated as gym environment in Python to prepare reinforcement learning training, the decisions of the fully trained system are then given to operators via classical visual description. The evaluation is performed in the digital twin, real-world implementation, including gaining operators trust and navigating more complex interdependencies, is ongoing.

4 REINFORCEMENT LEARNING APPROACH

Reinforcement learning (RL) presents a compelling approach for addressing the real-time hoist scheduling problem in electroplating production systems. The core principle involves training an intelligent agent to make sequential decisions regarding hoist movements for material flow through interaction with a simulated environment, here the digital twin, that accurately models the physical system's constraints and dynamics. Thus, the classical approach of regarding an RL agent as an entity that receives an observation of a state of the environment at time t, denoted as s_t, to select an (optimal) action a_t, is still applicable. To foster good decision making, the agent receives a reward r_{t+1} in the next time step $t+1$. The environment then changes to s_{t+1} and the entire cycle repeats (Kuhnle et al. 2021).

A key challenge is the holistic optimization and expert knowledge guided selection of the action space, state space and reward scheme. In this case of electroplating, learning is hindered through the indirect connection of decisions. In particular, only a small fraction of the time is influenced by the RL agent, yet it is getting rewarded for the entire period as visualized exemplarily in Figure 3. E.g. initiating a transport entails transport time and buffer times which are (partially) influenced, yet dwell times and lifting, draining etc. are added to the total time, and hence reward, but not directly controlled by the RL agent.

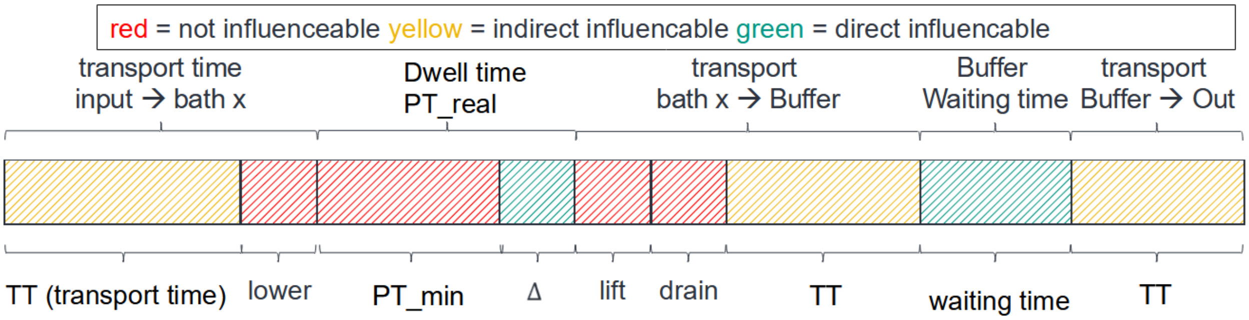
red = not influenceable yellow = indirect influencable green = direct influencable
transport time input → bath x
Dwell time PT_real
transport bath x → Buffer
Buffer Waiting time
transport Buffer → Out
TT (transport time)
lower
PT_min
Δ
lift
drain
TT
waiting time
TT

Figure 3: Influence on required times by the RL agent where Δ signifies $PT_{real} - PT_{min}$.

4.1 Methodology Framework

The RL framework for electroplating hoist scheduling comprises several interconnected components. Initially, the state space encompasses the current system configuration, including hoist positions, bath occupancy status, processing time windows of active orders, and criticality metrics based on maximum dwell time proximity. The action space consists of possible hoist movements and order transport operations, constrained by physical system limitations such as single-track operations and exclusive bath access zones.

The reward function, from a business perspective, should maximize the throughput which integrates multiple production objectives such as minimization of lead time, reduction of constraint violations, and optimization of resource utilization. However, due to the indirect control over the dwell time and hence the lead times and utilization by controlling the hoists, it is sufficient to regard the number of jobs performed per hour and the number of maximum dwell time violations. This study excludes potential further processes outside the system, which could also be influenced by the RL agent's decisions.

4.2 Training and Implementation Methodology

Agent training occurs within the digital twin environment, which simulates the stochastic nature of the electroplating system as foresight, including equipment failures, variable times, and manual interventions. Foresight refers to copying the digital twin at the present moment into an equally instantiated simulation with current randomization (May et al. 2021). This approach enables the agent to learn robust policies that accommodate system variability without requiring exhaustive enumeration of potential scenarios. The process is based on the digital twin and simulation through a SimPy DES and includes modules for state preparation based on the raw observation of the system through preprocessing, an action module that includes action masking to improve learning (Loffredo et al. 2023) and the reward calculation. The environment is based on the gym environment to be interchangeably connectable with common RL python packages. For this study, RLlib is used due to the support of Ray Tune for automated hyperparameter tuning in an attempt towards AutoRL (AutoML for RL). The methodology and framework is visualized in Figure 4. The digital twin itself was validated according to the operational performance validation approach (May et al. 2024) and is verified in the beginning of any foresight period (May et al. 2021).

Deep Q-Networks (DQN) or Proximal Policy Optimization (PPO) algorithms provide effective mechanisms for policy learning in high-dimensional state spaces (Kuhnle et al. 2021). Based on the initial experiments and typical superiority of PPO in handling instability in general and in manufacturing settings (Kuhnle et al. 2022), PPO is chosen in the RLlib package. The system is designed as a single agent and competes with the heuristic performance, which represents the industrial standard. However, in the course of the training procedures and the selection of the best combination of action, state and reward modeling, intermediate steps are performed. For instance, at first, a valid agent, which solely has to learn to perform valid actions, is trained. If this is not possible, likely a good process control cannot be achieved (Loffredo et al. 2023).

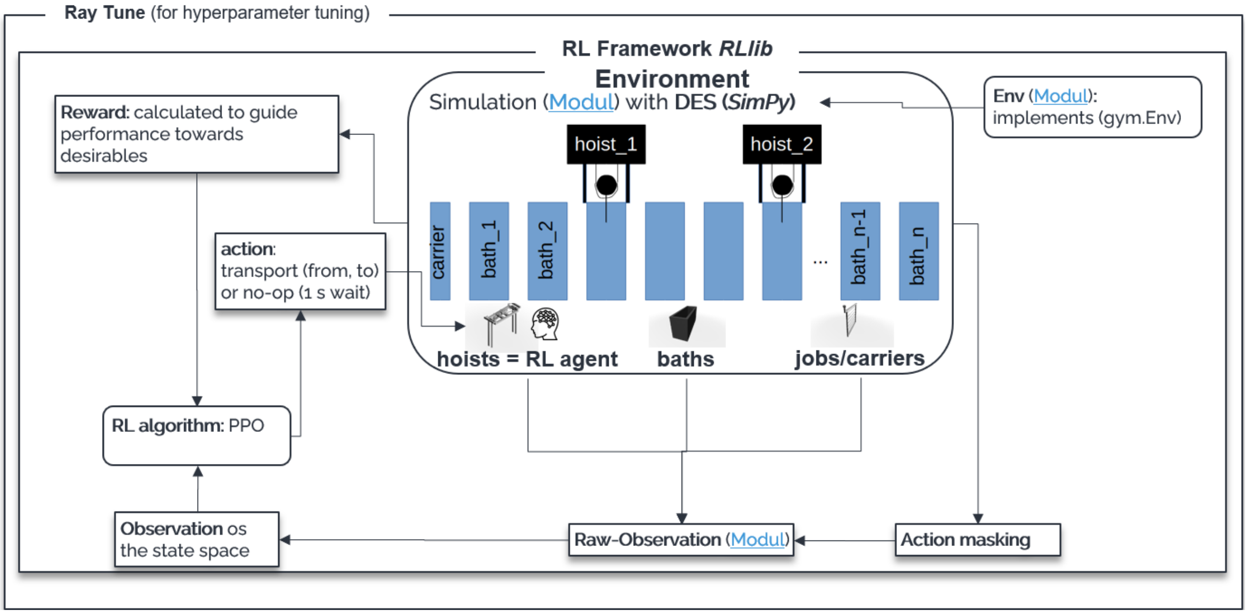

Figure 4: RL training framework.

4.3 State definition

The state representation constitutes a critical component as the representation must effectively encapsulate all relevant system information while maintaining computational efficiency. The implemented state space comprises a one-dimensional vector that concatenates six distinct informational components:

1. Position of Hoists: This vector encodes the spatial distribution of hoists across the system's baths. Each hoist's position is represented by a binary indicator (one) at the corresponding bath location within the vector. The vector length equals the total number of baths across all hoists' operational zones, with the number of nonzero entries corresponding to the number of hoists in the system.
2. State of Hoists: This component represents the operational status of each hoist through binary indicators, where one signifies availability and zero indicates engagement in a transport operation. The dimension of this vector equals the total number of hoists in the system.
3. State of Baths: Following analogous encoding principles, this vector represents bath occupancy status throughout the system. A value of one indicates an occupied bath, zero represents availability.
4. Duration of Empty Runs: This vector quantifies the temporal requirements for hoists to traverse from their current positions to any other bath within their operational zones. The current position of each hoist is marked with zero, while other entries contain integer values representing required transport times in seconds. Note that the loaded transport times may differ depending on the load.
5. Remaining Time to Minimum Dwell Time: This component tracks the remaining time until orders in occupied baths reach their minimum processing requirements. A value of zero may indicate either precise timing completion or bath vacancy, while negative values signify that minimum dwell time requirements have been satisfied and the order is eligible for transport.
6. Remaining Time to Maximum Dwell Time: This vector monitors the critical temporal constraints by indicating remaining time until maximum dwell time thresholds are exceeded for each bath. This information is essential for preventing quality degradation through overprocessing.

4.4 Action definition

The action space formulation encompasses a comprehensive set of decision options, while maintaining computational efficiency, with the action domain comprising both transport actions and strategic waiting operations. Transport actions are formalized as ordered pairs (S, D), where S represents the source bath and D indicates the destination bath or carrier loading and unloading bay. This representation implicitly incorporates multiple atomic operations: activation of bath S, extraction of the order from bath S, transportation, and subsequent immersion into bath D. For systems with overlapping operational zones between multiple hoists, this representation is extended to include the executing hoist κ, yielding the triple (κ, S, D).

To enhance computational efficiency and accelerate learning convergence, domain-specific knowledge is leveraged to constrain the action space. Only bath transitions that correspond to valid processing sequences within the operational programs are included in the action space. This a priori elimination of infeasible actions significantly reduces the dimensionality of the decision space without compromising operational functionality. Further refinement of the action space is achieved through implementation of action masking techniques. This approach dynamically filters actions based on current system states, rendering only feasible actions available for selection. An action is deemed feasible when its source bath currently contains an order available for transport. This dynamic constraint mechanism prevents the agent from attempting invalid operations and focuses the learning process on meaningful action sequences (Loffredo et al. 2023).

Complementing the transport actions, the action space incorporates waiting actions that provide temporal flexibility in decision-making. These waiting actions enable the agent to strategically delay transport operations when immediate action would be suboptimal, i.e. just before reaching a minimum dwell time in a neighboring bath. The implemented system offers three distinct waiting durations: one, five, and ten seconds. Upon selection of a waiting action, the system advances according to the specified duration, after which the agent reevaluates the new system state and selects a subsequent action based on updated conditions.

4.5 Reward definition

The composite reward function incorporates multiple order-specific components that address critical performance metrics while balancing potentially competing objectives. In short, it can be regarded as a simultaneous multi-objective real-time scheduling approach (Hofmann et al. 2022). First is a Lead Time Optimization (Component A): This reward term incentivizes minimization of order processing time by rewarding efficient decisions. The exponential formulation provides graduated reinforcement proportional to the improvement achieved relative to theoretical minimum processing times (Kuhnle et al. 2021):

$$R_A = e^{k \cdot \Delta t} \quad \forall \Delta t > 0$$

where $\Delta t = t_{\text{current task}} - t_{\text{theo\_min\_task type}}$ and $k = \frac{\ln(0.8)}{0.2 \cdot tpt_{\text{theo\_min\_task type}}}$. k can be determined in preliminary experiments during a pre-study.

Second is the Maximum Dwell Time Penalty (Component B): This penalty term discourages exceeding (by t') the maximum dwell time constraints (based on the costs ρ_i of the order i) by implementing progressively increasing negative reinforcement when orders exceed their maximum allowable processing times. Note that there is a lower bound if the maximum and minimum time are very close:

$$R_B = \sum_i \rho_i \cdot \left(e^{k \cdot t'} - 1 \right) \quad \forall t \geq t_{\max}$$

where $k = \frac{\ln(0.5/\rho_i)}{t_{\max} - t_{\min}}$ for $t_{\max} - t_{\min} > 10$ and $k = \frac{\ln(0.5/\rho_i)}{10}$ for $t_{\max} - t_{\min} \leq 10$.

The individual reward components are integrated into a comprehensive reward function that guides agent behavior toward optimizing system performance:

$$R_{\text{total}} = \alpha \cdot R_A - \beta \cdot R_B$$

where α and β are weighting coefficients that balance the relative importance of the respective objectives subject to α not significantly exceeding β. These coefficients can be tuned to emphasize particular operational priorities based on specific production requirements. The reward is given after each completion of an active decision. Non active decision that correspond to waiting are rewarded with constant 0.

5 EVALUATION

For validation purposes, the proposed approach is implemented and validated in the previously explained real-world galvano electroplating use-case. Digital twin creation is verified over three months and the training and evaluation scenario reaches the real-world equivalent of years. As visualized in Figure 5 the proposed RL agent within the digital twin framework can significantly outperform the industrial heuristics used. Through adaption of the state space, action space and most importantly the weighting of the reward function, specifically well performing RL policies can be derived. For instance, the visualized RL red is hyperparameter tuned after the reward function is set to enforce high throughput and creates a performance improvement of about 3%. Similarly, when focusing on not violating the maximum dwell times (PPO black), an performance increase of about 4% can be realized. However, even the existing heuristics perform very well, as they are positioned close to the pareto optimal PPO black. Thus, selection of preferences is required and only PPO black slightly dominates the heuristics, i.e. having lover maximum dwell time violations and higher finished parts per hour than the heuristics at the same time. Moreover, when comparing the performance with a random, average performing, non-hyperparameter tuned RL agent the importance of fine-tuning to the specific reward function becomes clear.

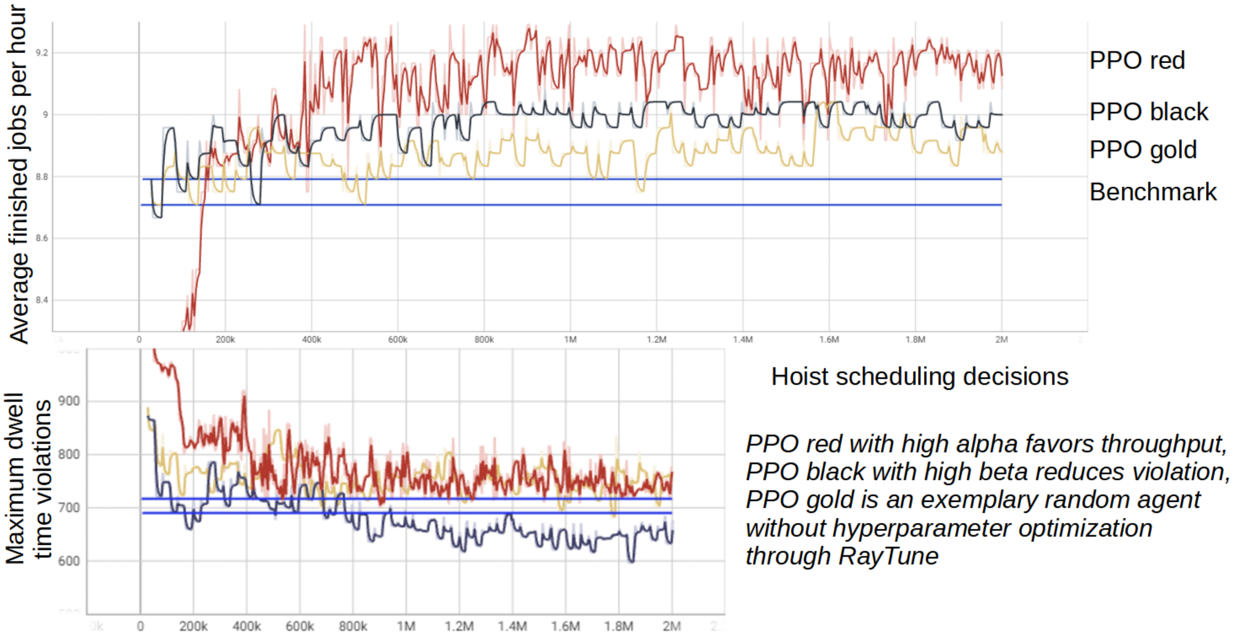

Figure 5: RL performance evaluation in the digital twin compared to the baseline industrial heuristic.

Thus, the two research questions about the digital twin framework (RQ2) and the general possibility to use RL to solve the hoist scheduling problem (RQ1) can be positively answered. However, there is still room for further improvement and studies, in particular as larger systems in electroplating are the norm and the proposed approach could be extended in style of Kuhnle et al. (2022) to derive a humanly understandble RL policy that satisfies the industrial need for simple, understandable heuristics (Reimschüssel et al. 2023).

6 CONCLUSION

The proposed reinforcement learning approach for hoist scheduling in electroplating production systems demonstrates significant potential for operational optimization within a digital twin framework. Through systematic adaptation of the state space, action space, and reward function weighting, performance improvements of 3-4% compared to existing industrial heuristics can been achieved in this case study. The reward function parameterization proves particularly influential, with specialized configurations enabling targeted optimization toward either lead time minimization or constraint violation minimization. Implementation in real-world electroplating facilities represents the logical next step in validating the approach under authentic production conditions with larger systems that comprise multiple hoists, tracks and potentially require a multi-agent system. Additionally, integration with existing manufacturing execution systems and development of user-friendly interfaces would facilitate practical adoption in industrial environments. Further investigation into the approach's robustness against stochastic disruptions and its adaptability to varying product mixes could enhance its applicability across diverse production scenarios. Last but not least, the early implementation of such a digital twin, not for the control purposes used herein, but also for production and layout planning could improve electroplating and the general manufacturing industry.

ACKNOWLEDGMENTS

The authors would like to acknowledge the discussions, data provision and support from industrial partners.

REFERENCES

Beiner, S., M. Kandler, D. Richter, M. C. May, S. Kinkel, and G. Lanza. 2023. “Artificial Intelligence Implementation Strategy for Industrial Companies Using the AI Tool Box - A Morphology for Selecting Relevant AI Use Cases”. *Lecture Notes in Mechanical Engineering*:763–773.

Chauvet, F., E. Levner, L. K. Meyzin, and J.-M. Proth. 2000. “On-line Scheduling in a Surface Treatment System”. *European Journal of Operational Research* 120(2):382–392.

Chen, T., V. Sampath, M. C. May, S. Shan, O. J. Jorg, J. J. Aguilar Martín, *et al*. 2023. “Machine Learning in Manufacturing towards Industry 4.0: From ‘For Now’to ‘Four-Know”’. *Applied Sciences* 13(3):1903.

Chen, X., B. Yang, Z. Pang, P. Zhou, and G. Fu. 2024. “Coupled Optimization of Task Sequence and Hoist Scheduling for Electroplating Production Lines based on an improved SALP Swarm Algorithm”. *CIRP Journal of Manufacturing Science and Technology* 53:34–47.

Chové, E., P. Castagna, and R. Abbou. 2009. “Hoist Scheduling Problem: Coupling reactive and predictive approaches”. *IFAC Proceedings Volumes* 42(4):2077–2082.

Chtourou, S., M.-A. Manier, and T. Loukil. 2013. “A hybrid algorithm for the Cyclic Hoist Scheduling Problem with two Transportation Resources”. *Computers & Industrial Engineering* 65(3):426–437.

Hofmann, C., X. Liu, M. May, and G. Lanza. 2022. “Hybrid Monte Carlo Tree Search based Multi-Objective Scheduling”. *Production Engineering* 17(1):133–144.

Kuhnle, A., J.-P. Kaiser, F. Theiß, N. Stricker, and G. Lanza. 2021. “Designing an Adaptive Production Control System using Reinforcement Learning”. *Journal of Intelligent Manufacturing* 32:855–876.

Kuhnle, A., M. C. May, L. Schäfer, and G. Lanza. 2022. “Explainable Reinforcement Rearning in Production Control of Job Shop Manufacturing System”. *International Journal of Production Research* 60(19):5812–5834.

Lee, J.-H., and H.-J. Kim. 2022. “Reinforcement Learning for Robotic Flow Shop Scheduling with Processing Time Variations”. *International Journal of Production Research* 60(7):2346–2368.

Leiden, A., C. Herrmann, and S. Thiede. 2021. “Cyber-physical Production System Approach for Energy and Resource Efficient Planning and Operation of Plating Process Chains”. *Journal of Cleaner Production* 280:125160.

Loffredo, A., M. C. May, A. Matta, and G. Lanza. 2024. “Reinforcement Learning for Sustainability Enhancement of Production Lines”. *Journal of Intelligent Manufacturing* 35(8):3775–3791.

Loffredo, A., M. C. May, L. Schäfer, A. Matta, and G. Lanza. 2023. “Reinforcement Learning for Energy-efficient Control of Parallel and Identical Machines”. *CIRP Journal of Manufacturing Science and Technology* 44:91–103.

Lugaresi, G., S. Gangemi, G. Gazzoni, and A. Matta. 2022. “Online Validation of Simulation-based Digital Twins Exploiting Time Series Analysis”. In *2022 Winter Simulation Conference (WSC)*, 2912–2923 https://doi.org/10.1109/WSC57314.2022.10015346.

Lugaresi, G., and A. Matta. 2018. “Real-time Simulation in Manufacturing Systems: Challenges and Research Directions”. In *2018 Winter Simulation Conference (WSC)*, 3319–3330 https://doi.org/10.1109/WSC.2018.8632542.

Manier, M.-A., and C. Bloch. 2003. “A classification for hoist scheduling problems”. *International Journal of Flexible Manufacturing Systems* 15:37–55.

Manier, M.-A., and S. Lamrous. 2008. “An Evolutionary Approach for the Design and Scheduling of Electroplating Facilities”. *Journal of Mathematical Modelling and Algorithms* 7:197–215.

May, M. C., A. Albers, M. D. Fischer, F. Mayerhofer, L. Schäfer, and G. Lanza. 2021. “Queue Length Forecasting in Complex Manufacturing Job Shops”. *Forecasting* 3(2):322–338.

May, M. C., L. Kiefer, and G. Lanza. 2024. “Digital Twin Based Uncertainty Informed Time Constraint Control in Semiconductor Manufacturing”. In *2024 Winter Simulation Conference (WSC)*, 1943–1954 https://doi.org/10.1109/WSC63780.2024.10838845.

May, M. C., C. Nestroy, L. Overbeck, and G. Lanza. 2024. “Automated Model Generation Framework for Material Flow Simulations of Production Systems”. *International Journal of Production Research* 62(1-2):141–156.

May, M. C., J. Oberst, and G. Lanza. 2024. “Managing Product-inherent Constraints with Artificial Intelligence: Production Control for Time Constraints in Semiconductor Manufacturing”. *Journal of Intelligent Manufacturing*:1–18.

May, M. C., L. Overbeck, M. Wurster, A. Kuhnle, and G. Lanza. 2021. “Foresighted Digital Twin for Situational Agent Selection in Production Control”. *Procedia CIRP* 99:27–32.

May, M. C., L. Schäfer, A. Frey, C. Krahe, and G. Lanza. 2023. “Towards Product-Production-CoDesign for the Production of the Future”. *Procedia CIRP* 119:944–949.

Overbeck, L., V. Glaser, M. C. May, and G. Lanza. 2023. “Generalization of Reinforcement Learning Agents for Production Control”. *Lecture Notes in Mechanical Engineering*:338–346.

Overbeck, L., S. C. Graves, and G. Lanza. 2024. “Development and Analysis of Digital Twins of Production Systems”. *International Journal of Production Research* 62(10):3544–3558.

Overbeck, L., A. Hugues, M. C. May, A. Kuhnle, and G. Lanza. 2021. “Reinforcement Learning Based Production Control of Semi-automated Manufacturing Systems”. *Procedia CIRP* 103:170–175.

Paul, H. J., C. Bierwirth, and H. Kopfer. 2007. “A Heuristic Scheduling Procedure for Multi-item Hoist Production Lines”. *International Journal of Production Economics* 105(1):54–69.

Ptuskin, A., E. Levner, and V. Kats. 2024. “Cyclic Multi-hoist Scheduling with Fuzzy Processing Times in Flexible Manufacturing Lines”. *Applied Soft Computing* 165:112014.

Reimschüssel, S., U. Fuchs, and G. Sand. 2023. “Electroplating Scheduling: Closing a Research Gap from an Automation Vendor’s Perspective”. In *Computer Aided Chemical Engineering*, Volume 52, 125–130. Elsevier.

Subaï, C., P. Baptiste, and E. Niel. 2006. “Scheduling Issues for Environmentally Responsible Manufacturing: The case of Hoist Scheduling in an Electroplating Line”. *International Journal of Production Economics* 99(1-2):74–87.

Wurster, M., M. Michel, M. C. May, A. Kuhnle, N. Stricker, and G. Lanza. 2022. “Modelling and Condition-based Control of a Flexible and Hybrid Disassembly System with Manual and Autonomous Workstations using Reinforcement Learning”. *Journal of Intelligent Manufacturing* 33(2):575–591.

Xiao, Y., K. Jin, R. Ma, and H. H. Zhuo. 2024. “Large-Scale Electroplating Scheduling: A Hierarchical Temporal Planning Approach”. In *Proceedings of the International Conference on Intelligent Computing, August 5th, Tianjin, China*, 215–226.

Yih, Y. 1992. “Learning Real-time Scheduling Rules from Optimal Policy of Semi-Markov Decision Processes”. *International Journal of Computer Integrated Manufacturing* 5(3):171–181.

Yih, Y., T.-P. Liang, and H. Moskowitz. 1993. “Robot Scheduling in a Circuit Board Production Line: A Hybrid OR/ANN Approach”. *IIE Transactions* 25(2):26–33.

Yih, Y., and A. Thesen. 1991. “Semi-Markov Decision Models for Real-time Scheduling”. *International Journal of Production Research* 29(11):2331–2346.

AUTHOR BIOGRAPHIES

MARVIN CARL MAY is an Assistant Professor for Industrial AI at Nanyang Technological University (NTU) in Singapore and was a postdoctoral researcher at MIT. His research interests include Production Planning and Control, Product-Production-CoDesign, Simulation based optimization and Machine Learning. His email addresses are mc_may@mit.edu, marvin.may@ntu.edu.sg.

LOUIS SCHÄFER is a researcher at KIT and adesso SE. His research interests include Production Planning, Simulations and Machine Learning for control within manufacturing. His email address is louis.schaefer@adesso.de.

JAN-PHILIPP KAISER is a postdoctoral researcher at KIT. His research interests include Remanufacturing, Planning of complex manufacturing and Machine Learning for control within manufacturing. His email address is jan-philipp.kaiser@kit.edu.